SSIG and IRISA at Multimodal Person Discovery

Cassio E. dos Santos Jr[1], Guillaume Gravier[2], William Robson Schwartz[1]
[1]Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[2]IRISA & Inria Rennes , CNRS, Rennes, France
cass@dcc.ufmg.br, guig@irisa.fr, william@dcc.ufmg.br

ABSTRACT

This paper describes our approach and results in the multimodal person discovery in broadcast TV task at MediaEval 2015. We investigate two distinct aspects of multimodal person discovery. One refers to face clusters, which are considered to propagate names associated to faces in one shot to other faces that probably belong to the same person. The face clustering approach consists in calculating face similarities using partial least squares (PLS) and a simple hierarchical approach. The other aspect refers to tag propagation in a graph-based approach where nodes are speaking faces and edges link similar faces/speakers. The advantage of the graph-based tag propagation is to not rely on face/speaker clustering, which we believe can be errorprone.

1. INTRODUCTION

Multimodal person discovery in video archives consists in naming all speaking faces in the collection without prior information, leveraging face recognition, speech recognition, speaker recognition and optical character recognition. A description of the task and resources provided within MediaEval is given in [2]. In particular, two key components of most systems for multimodal person discovery are (i) face tracking and clustering and (ii) speaker diarization. See [6] for a recent overview of existing systems. Given these components, a popular strategy to name speaking faces relies on a mapping of face clusters and speakers from the diarization, combining this mapping with appearance of named entities in speech transcripts or on screen (e.g., [3, 8]). The baseline system provided by the organizers [7] is a clear instanciation of this. Person names appearing on screen are first propagated onto speaker clusters, finding an optimal mapping based on co-occurrence. In the next step, one has to find for each named speaker if there is a co-occurring face track that has a probability to correspond to the current speaker higher than a threshold. Each such face track receives the name assigned to the speaker cluster.

We explore two distinct aspects of multimodal person discovery in this evaluation. On the one hand, we seek to improve face clustering using recent advances in face recognition based on partial least square (PLS) regression [4]. We consider a variant of the baseline system provided, modified to better merge the PLS face cluster and speaker diarization results. On the other hand, we study tag propagation in a graph where nodes are speaking faces, with edges denoting the voice and/or face similarity. This approach is motivated by the wish to avoid explicit face and speaker clustering and open new strategies for person discovery. Note that the two approaches could be combined but, for practical reasons, this combination was not considered in the framework of the evaluation.

2. PLS-BASED FACE CLUSTERING

The PLS-based face clustering approach consists in calculating a similarity measure between face tracks for further clustering. Face clusters are then used in a variant of the baseline, as a replacement of the face clusters provided.

PLS is a statistical method consisting of two steps: regression and projection [9]. The projection step consists in calculating a subspace that maximize the covariance between predictors and responses. The regression step relies on ordinary least squares to estimate responses based on the projected predictors. We employ the one-shot similarity metric based on PLS for face verification described in [4], which presents robust results for face images in the wild compared to conventional distance-based methods. In a nutshell, the similarity $sim(A, B)$ between face tracks A and B relies on PLS regression trained to return $+1$ for samples in A and response -1 for samples in a background set of images (300 random face images from the LFW dataset [5]). Then, $sim(A, B)$ is calculated as the average of responses from samples in B evaluated in the learned PLS regression. A symmetric version is used in practice, averaging $sim(A, B)$ and $sim(B, A)$.

Based on PLS similarity calculated between all face track pairs, clustering aims at grouping face tracks from the same subject. We employ a hierarchical clustering approach that consists in merging a pair of face tracks with maximum similarity and with at least one face track that was not merged yet. The merging consists in propagating an identification label from one face track to the other or generating a new identification label for the pair if no label was previous associated to the face tracks. The algorithm stops when the maximum similarity is less than a threshold, empirically set to 0.5 using the development set.

To assess the interest of PLS-based face clustering, we consider a slightly different version of the baseline approach to merge face clustering and speaker diarization information. Each name associated to one face track is propagated to all face tracks within the same face cluster. We then consider the union of the names from the face tracks and speaker diarization within each shot. We also evaluate the

Copyright is held by the author/owner(s).
MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

method	BSLN	SPKR	FACE	UNI	INT
dev	38.89	63.67	49.12	67.84	44.83
test	78.35	89.46	67.18	**89.74**	66.86
test (PLS)	78.35	89.46	61.90	89.64	61.64

Table 1: EwMAP (in %) using the baseline face clusters on the development set (top row), on the test set (middle row) and using the PLS-based face clusters on the test set (bottom row).

modified baseline approach using only the speaker diarization, only the face cluster, and considering the intersection of the names instead of union.

3. GRAPH-BASED TAG PROPAGATION

To skirt issues with errors in clustering, which we believe can strongly affect the naming process, we investigate a strategy based on tag propagation within a graph where a node corresponds to an occurrence of a speaking face within a shot.

The first step is the graph construction process, which consists in identifying speaking faces from the face tracks detected within each shot[1]. This is achieved by selecting face tracks whose probability to correspond to the current speech turn is greater than a threshold empirically set to 0.6, where the probabilities that a face track corresponds to a speech turn are those provided. For each selected face track, we keep a record of the matching speech turn. The selected speaking face tracks are the nodes of a graph and are connected with edges bearing two scores, depicting the similarity of resp. voice and face (as given in the speech turn and face track similarity files). To avoid a fully connected graph and keep only relevant relationships, we connect two nodes if the similarity between the corresponding face tracks and the similarity between the corresponding speech turns are both above a threshold, empirically set to 0.1 for both modalities. Note that having no relations between face tracks and speech turns across shows, a graph is built independently for each show.

The naming process starts by associating a name to a node whenever possible based on the output of overlaid text detection: if an overlay significantly overlaps the face track, the node is tagged with the corresponding name and a score of 1. In case of multiple overlaping overlays, the name corresponding to the longest co-occurrence is considered. After tagging all nodes, tags are optionally propagated over a number of iterations. At each iteration, each tag of each node is propagated via the corresponding edges with a propagation score equal to the tag score multiplied by the edge weight, where edge weights are taken as the average of the face and voice similarity. After propagation, each node receives the tag with the highest score.

4. RESULTS

The results from the second submission (July 8th) of the four PLS-based methods and the baseline are presented in Tab. 1, where the following abbreviations are employed: PLS-based face clustering considering only speaker diarization (SPKR), only face clusters (FACE), union (UNI) and intersection (INT) of names among face clusters and speaker diarization. In PLS-based face clustering, we consider the CLBP [1] feature descriptor with radius parameter 5 calculated in squared blocks of size 16 pixels and stride of 8 pixels. All faces were cropped from the videos using the face position provided in the baseline approach and scaled to 128 by 128 pixels. Note that we do not provide face clusters based on PLS for the development set and, therefore, all results in Tab. 1 for the development set consider only the face clusters available in the baseline approach. We also provide the results on the test set considering the face clusters provided in the baseline method, i.e., without PLS-based face clustering.

The SPK approach yields the best EwMAP in Tab. 1 while the FACE yields the worst results. However, the results from INT and UNI indicate that the two approaches present complementary results, i.e., the intersection of the propagated names among the face clusters and speaker diarization shots indicates that a small subset of correct names from the face clusters that are not in the speaker names, These aspects are observed in the development and test set, using the face clusters in the baseline or the PLS method face clusters. We also noticed no significant difference in the results between face clusters provided in the baseline approach and using the PLS-based method, considering the UNI approach. We believe that this small difference is an effect of the poor quality of the face clusters, which might result from combined errors in the face detection and in the face tracking methods.

		EwMAP	MAP	C
dev	no prop	44.5	44.7	76.7
	1 step prop	53.6	54.0	75.4
test	no prop	78.3	79.5	89.7

Table 2: Results with graph-based naming on the development data (test2) and on the test data.

Results for the graph-based tag propagation method are given in Tab. 2. On the development data (test2 subset), results are provided without tag propagation (no prop) and with a singl step of tag propagation. We believe that the poor results obtained are attributable to the fact that the graph links only submission shots, which account only for a small fraction of the total number of shots in the development data. Contrarily, most of the shots in the test data are subission shots. With no surprise, tag propagation improves the MAP to the expense of correctness. Submission on the test set was made without tag propagation (because of unconvincing propagation results at the time) and not updated after the initial submission (July 1st). Interestingly, direct naming of speaking face tracks from overlays (i.e., no propagation) already provides accurate tagging.

5. ACKNOWLEDGEMENTS

The authors would like to thank the Brazilian National Research Council – CNPq (Grant #477457/2013-4), Brazilian National Council for the Improvement of Higher Education – CAPES (Grant STIC-AMSUD 001/2013) and the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-01806-13 and CEX-APQ-03195-13). This work was partially supported by the STIC AmSud program, under the project 'Unsupervised Mining of Multimedia Content', and by the Inria Associate Team program.

[1] Only submission shots were considered in this work.

6. REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] H. Bredin, J. Poignant, and C. Barras. Overview of the multimodal person discovery task at MediaEval 2015. In *Working Notes Proc. of MediaEval 2015 Workshop*, 2015.

[3] H. Bredin, A. Roy, V.-B. Le, and C. Barras. Person Instance Graphs for Mono-, Cross- and Multi-Modal Person Recognition in Multimedia Data. Application to Speaker Identification in TV Broadcast. *International Journal of Multimedia Information Retrieval*, 2014.

[4] H. Guo, W. R. Schwartz, and L. S. Davis. Face verification using large feature sets and one shot similarity. In *Intl. Conf. on Biometrics*, pages 1–8, 2011.

[5] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[6] J. Poignant. *Identification non-supervisée de personnes dans les flux teéleévisés.* PhD thesis, Université de Grenoble, 2013.

[7] J. Poignant, H. Bredin, V.-B. Le, L. Besacier, C. Barras, and G. Quénot. Unsupervised speaker identification using overlaid texts in TV broadcast. In *Annual Conf. of the International Speech Communication Association*, 2012.

[8] J. Poignant, G. Fortier, L. Besacier, and G. Quénot. Naming multi-modal clusters to identify persons in TV broadcast. *Multimedia Tools and Applications*, 2015.

[9] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Subspace, latent structure and feature selection*, pages 34–51. Springer, 2006.